ACTIS: Improving data efficiency by leveraging semi-supervised Augmentation Consistency Training for Instance Segmentation

Josef Lorenz Rumberger[1,2,3,5] Jannik Franzen[1,2,5] Peter Hirsch[1,3,5]
Jan-Philipp Albrecht[1,3,5] Dagmar Kainmueller[1,4,5]
[1] Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany
[2] Charité University Medicine, Berlin, Germany
[3] Humboldt-Universität zu Berlin, Faculty of Mathematics and Natural Sciences, Berlin, Germany
[4] Potsdam University, Digital Engineering Faculty, Germany
[5] Helmholtz Imaging firstnames.lastname@mdc-berlin.de

Abstract

*Segmenting objects like cells or nuclei in biomedical microscopy data is a standard task required for many downstream analyses. However, existing pre-trained models are continuously challenged by ever-evolving experimental setups and imaging platforms. On the other hand, training new models still requires a considerable number of annotated samples, rendering it infeasible for small to midsized experiments. To address this challenge, we propose a semi-supervised learning approach for instance segmentation that leverages a small number of annotated samples together with a larger number of unannotated samples. Our pipeline, **A**ugmentation **C**onsistency **T**raining for **I**nstance **S**egmentation (ACTIS), incorporates methods from consistency regularization and entropy minimization. In addition, we introduce a robust confidence-based loss masking scheme which we find empirically to work well on highly imbalanced class frequencies. We show that our model can surpass the performance of supervised models trained on more than twice as much annotated data. It achieves state-of-the-art results on three benchmark datasets in the biomedical domain, demonstrating its effectiveness for semi-supervised instance segmentation. Code: https://github.com/Kainmueller-Lab/ACTIS*

1. Introduction

Increasing amounts of microscopy data in the biomedical domain have necessitated the development of automated analysis methods. An essential step prior to many downstream analyses is the segmentation of objects, often cells or nuclei, within the images. A number of popular pre-trained models for cell [15] and nuclei segmentation [39, 36] already exist, but they may not generalize to a novel sample preparation protocol, a novel imaging modality, or a specific cell population if these are not represented in the respective training data [10].

Fine-tuning a pre-trained model [21, 37] can greatly improve annotation efficiency compared to training a new model from scratch: If the domain-gap between the pretraining- and fine-tuning datasets is sufficiently small [18], it can substantially reduce the respective time and cost requirements [15]. For larger domain-gaps, self-supervised pre-training, e.g., via contrastive-learning [37] or generative pre-training [42], can be employed to obtain a good initialization for task-specific fine-tuning [20]. However, self-supervised pre-training typically requires large datasets, substantial computational resources, and additional engineering.

To this end, semi-supervised learning presents a popular alternative: By leveraging a small number of annotated samples together with a larger number of non-annotated samples in a joint training objective, semi-supervised approaches can achieve comparable performance to models trained fully-supervised on much larger annotated datasets [25, 4, 44]. Notable early works focused on minimizing the entropy of the prediction on the unlabeled samples [14, 25, 3]. More recent works add augmentation consistency regularization, i.e., a teacher model predicts pseudo-labels with weak augmentations and a student model predicts on inputs with strong augmentations, then the student is optimized for consistency with the teacher's predictions [41, 2, 47, 48].

However, most works on semi-supervised learning for image analysis focus on image classification tasks [47, 48, 41, 25, 2], a smaller number on semantic segmentation [38, 3, 31, 1, 46] and only very few on instance segmentation [6, 4, 45, 17], Within the latter, only two approaches [45, 17] employ consistency regularization and entropy minimization. Both approaches are *proposal-based*, i.e., they assume that bounding-boxes suitably approximate object instance shapes. This assumption is often violated in the biomedical domain, where objects may span large parts of an image and form dense clusters. This entails wide popularity of proposal-free methods for biomedical instance segmentation [15, 39]. However, to date, this class of methods has not been studied in conjunction with the predominant paradigms for semi-supervised learning, consistency regularization and entropy minimization.

To fill this gap, we here propose (1) a semi-supervised proposal-free instance segmentation pipeline that incorporates consistency regularization and entropy minimization. In addition, we introduce (2) a confidence-based loss masking scheme which we find empirically to work well for imbalanced class frequencies, and (3) we assess the performance of our model on three benchmark datasets from the biomedical domain, where we achieve state-of-the-art results compared to baseline semi-supervised and fully-supervised instance segmentation models.

2. Related Work

Entropy Minimization: Early works on semi-supervised learning focused on minimizing the entropy of the prediction on unlabeled samples. This can be done explicitly by adding an entropy regularization term to the objective function [14, 46] or implicitly by training against class-assigned predictions from the network, i.e., hard pseudo-labels, for unlabeled samples [25]. More recent works modify the latter approach by sharpening the predictions (i.e. concentrating the prediction around their maximum) instead of assigning classes [47, 38, 3], resulting in so-called soft pseudo-labels. A recent simplification of this approach [38] resorts to hard pseudo-labels, yet filters out samples from the loss calculation in case the maximum of the predicted confidence scores prior to class-assignment is below a certain threshold. We build upon this simplified approach [38], with a slight extension to tackle highly imbalanced class-frequencies: In this case, a class-agnostic fixed confidence threshold may lead to imbalanced filtering of the minority classes, since their confidence is typically lower than the one of more common classes. To avoid this, our method uses class-specific percentile-based confidence thresholds together with hard pseudo-labels, so that for every class the same percentage of high-confidence pixels is included in the loss calculation.

Augmentation Consistency Regularization: Early methods enforced consistency of predictions for different views (i.e., augmentations) of the same input data, as well as for different predictions obtained via stochastic neural network weights (e.g. via Dropout), by optimizing the mean-squared-error between different predictions [35]. This method has been extended by incorporating a temporal ensembling approach, where the average of multiple predictions from different views of the data at multiple time points during training is used as a pseudo-label [23]. The drawback of this method is the high memory footprint associated with storing multiple predictions for each unlabeled sample during training. This was addressed by [41], who introduce a teacher model with the same architecture as the student model, which constructs pseudo-labels on the fly during training. Instead of averaging over predictions, this approach averages over the student weights by updating the teacher weights as an exponential-moving-average of the students weights. It was found empirically to improve stability of pseudo-labels over the course of training and increase performance of the resulting student models significantly [41]. However, [3] show that the quality of the pseudo-labels can be further improved by averaging the predictions of multiple weakly augmented views of the data. Therefore, we use this method in conjunction with a momentum teacher to construct high-quality pseudo-labels [41, 3].

Contrastive Learning: Contrastive learning extends augmentation consistency regularization by constructing positive and negative pairs from different samples [7] or from patches within samples [46]. Positive pairs are constructed in the vain of augmentation consistency regularization, i.e., different views of the same sample, whereas negative pairs are constructed by taking different samples from a batch. Contrastive learning objectives then enforce consistent representations for the positive pairs and inconsistent representations for the negative pairs. However, many highly competitive semi-supervised learning approaches [38, 3, 2, 47] do not use contrastive learning and based on this observation, we did not include it into ACTIS.

Semi-Supervised Instance Segmentation To the best of our knowledge, the only other proposal-free semi-supervised instance segmentation method is Denoiseg [4], which uses self-supervised image denoising as an auxiliary task on the un-labeled samples and standard supervised training on labeled samples to deliver considerable improvements in segmentation quality. The datasets from this publication will be used for evaluating our method and for benchmarking it against their approach.

3. Method

We propose a competitive baseline model and a semi-supervised learning framework to further improve the model performance. The baseline model consists of a U-Net [32] with an ImageNet [9] pre-trained EfficientNet-B5 [40] backbone, which we train to predict foreground, background and instance boundary, i.e., a three-label instance segmentation model as in [5]. During test-time, the predictions are post-processed into individual segments via watershed transformation [8].

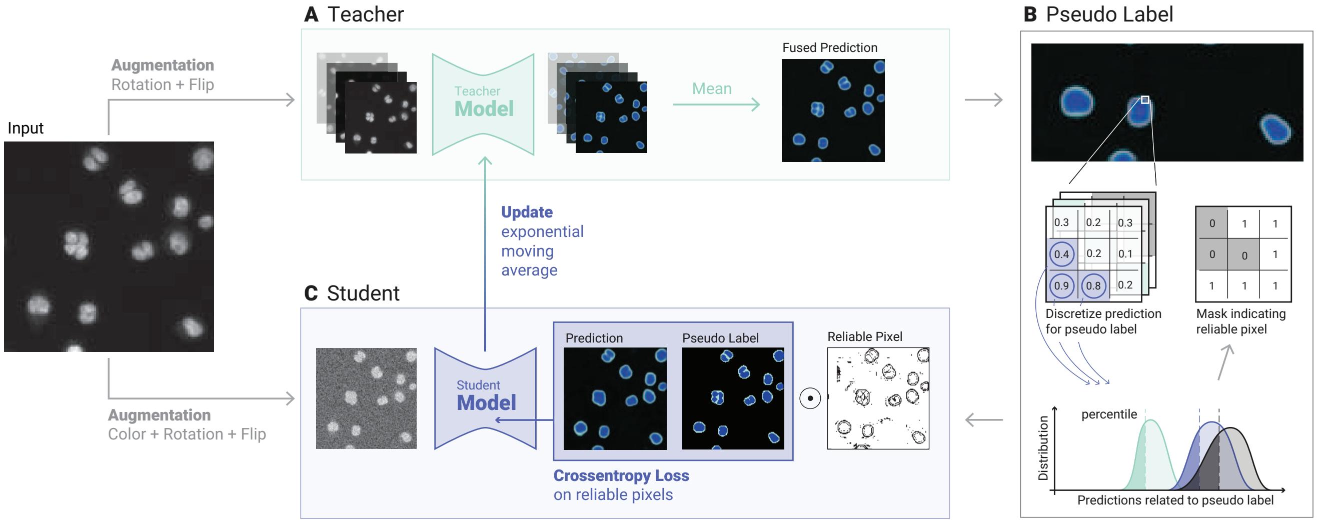

Figure 1. **Semi-supervised model architecture:** In A) the teacher produces a high-quality prediction by averaging over the predictions corresponding to multiple flipped and rotated versions of the input image. The prediction consists of 3 classes: background (black), cell boundary (turquoise), and cell interior (blue). In B) we discretize the resulting prediction to yield a 3-class pseudo-label. In addition to the pseudo-label, we construct a high-confidence mask by class-wise filtering of unreliable pixels. The pseudo-label and high-confidence mask are then used in C) to train a student model which is given the same but strongly augmented input.

The semi-supervised learning pipeline uses a Mean-Teacher [41] approach, i.e., the weights of the teacher are an exponential-moving-average of the student weights and both share the same architecture. Student and teacher are initialized at the start with the weights of a converged baseline model, which was trained fully-supervised on the same small labeled dataset that is used for supervision during semi-supervised training. Pseudo labels are generated by applying weak augmentations (flips, 90-degree rotations) on the teacher's input data and pixel-wise averaging of the outputs after respective inverse transforms as in [41]. Pixels within the pseudo-labels with predicted low confidence scores are considered unreliable and are thus excluded from the loss. The student is optimized by calculating the cross-entropy loss for the student's predictions of strongly augmented samples and the class-assigned and filtered pseudo-labels, together with regular supervised training on a small number of annotated samples. A graphical depiction of the semi-supervised training pipeline is given in Fig. 1.
The remainder of this section is organized as follows: Sec. 3.1 describes the confidence based filtering method, Sec. 3.2 describes the training pipeline, and Sec. 3.3 describes the post-processing.

3.1. Semi-supervised augmentation consistency training

Let us explore in more depth how we can make use of the large unlabeled dataset D in addition to the smaller labeled dataset D_l on which we pre-trained our baseline model:

$$D = \{x_i\}_{i=0}^{|D|}, \quad D_l = \{(x_i^l, y_i^l)\}_{i=0}^{|D_l|}$$

Here, x_i and x_i^l denote the unlabeled and labeled images, respectively and y_i^l the corresponding label. As described above, our model architecture consists of a teacher model f_{θ_t} and a student model f_{θ_s}, which initially are both copies of our pre-trained baseline. How to leverage D to improve the student model f_{θ_s} beyond the performance of the teacher model f_{θ_t} can be broken down into three parts as depicted in Fig. 1:
In **part A**, the teacher model is used to produce a high-quality prediction. To this end, multiple transformed versions $t_i(x)$ of an image x are taken as the input of the teacher model. As transformations $t_i(x) := (f_{\text{flip}} \circ f_{\text{rotate}})(x)$, we use rotations and flips but no affine transformations, because we want to avoid the necessity of using interpolations on the pseudo labels that might blur it. To yield the best quality prediction $\hat{y}$ we average over all q resulting predictions after applying the inverse transformation t_i^{-1}:

$$\hat{y} = \frac{1}{q} \sum_{i=0}^{q} t_i^{-1}(f_{\theta_t}(t_i(x))) \tag{1}$$

In **part B**, we use the high-quality prediction $\hat{y}$ to generate a discretized pseudo-label γ with elements

$$\gamma_j = \text{argmax}\{\hat{y}_{j,c} : c = 0...C\} \tag{2}$$

by assigning every pixel j the class c with the highest predicted confidence score. To filter out unreliable pixels in the pseudo-labels that might impair the supervision of the student, we construct a matrix M with pixel-wise elements

$$m_j = \begin{cases} 1 & \hat{y}_{j,c} > \bar{p}_c \text{ with } c = \gamma_j \\ 0 & \hat{y}_{j,c} \leq \bar{p}_c \text{ with } c = \gamma_j \end{cases} \tag{3}$$

which indicates a reliable pixel: A pixel is classified as reliable if the score $\hat{y}_{j,c}$ of the predicted class $c = \gamma_j$ surpasses the class-specific threshold $\bar{p}_c$. This threshold is constructed as follows. First, we compute for every class c the k-percentile p_c (here $k = 0.2$) of all teacher confidence scores $\hat{y}_{j,c}$ corresponding to the predicted class $c = \gamma_j$ in the pseudo label. In Fig. 1B, these scores are highlighted by the class-specific color. As the second step, we compute the exponential moving average

$$\bar{p}_c := p_{c,i+1} = 0.99 \cdot p_{c,i} + 0.01 \cdot p_c \tag{4}$$

of that percentile p_c over each training iteration i. Filtering based on individual percentiles per class ensures that rare classes like the boundary, which often have lower predicted confidence scores, are preserved during training.

Lastly in **part C**, the student receives the same input image x as the teacher but perturbed with strong intensity augmentations $t_{\text{int}}(x) := (f_{\text{color}} \circ f_{\text{jitter}} \circ f_{\text{blur}} \circ f_{\text{noise}})(x)$ together with flips and rotations t. The student model is trained to be robust against those perturbations. Hence, for an unlabeled sample $x \in D$ we use a cross-entropy loss

$$\mathcal{L}_{\text{semi}} = \mathcal{L}_{CE}\Big[f_{\theta_s}\big(t(t^{\text{int}}(x))\big), t(\gamma)\Big] \odot t(M). \tag{5}$$

By element-wise multiplication $\odot$ with the same spatially transformed matrix M, we filter out the unreliable pixels of the pseudo-label. In the case of a labeled sample $x^l \in D^l$ we use the standard cross-entropy loss

$$\mathcal{L}_{\text{supervised}} = \mathcal{L}_{CE}\Big[f_{\theta_s}\big(t(t^{\text{int}}(x^l))\big), t(y^l)\Big]. \tag{6}$$

Each training batch consists of labeled and unlabeled images. Thus, the total loss is the sum of the respective losses. In conclusion, only the weights of the student are optimized instantaneously towards augmentation consistency and entropy-minimization by minimizing the loss with respect to the weights $\min_{\theta_s} L$ via gradient descent. Whereas the teacher weights are updated as an exponential moving average of the student weights by applying the following formular $\theta_{t,n+1} = 0.99 \cdot \theta_{t,n} + 0.01 \cdot \theta_{s,n}$ once every 100 steps.

3.2. Training Details

The baseline model is a U-Net with EfficientNet-B5 encoder initialized with imagenet pre-trained weights. It optimizes a weighted cross-entropy loss for pixel-wise classification into background, foreground and boundary with class weights $(1, 1, 4)$ [5]. The baseline models are optimized with the AdamW optimizer and a cosine decay learning rate scheduler with a tuned learning rate, and relatively high weight decay of $1e^{-3}$ [26], which are best-practices known to yield highly competitive models [30, 33]. The data augmentation pipeline consists of flips, affine transformations, elastic deformations, gaussian blurring, additive gaussian noise and color jitter. Training data is split into 128×128 sized tiles as in Denoiseg [4], baseline models are trained with batchsize 5, whereas semi-supervised models are trained with batches consisting of 4 labeled and 8 unlabeled samples. Baseline models are optimized for 100,000 steps and the checkpoint with the highest AP$_{.50}$ score on the validation dataset is picked. Overfitting typically occurs early, especially for setups with a small number of labeled samples, so that all baseline models overfit or converge within the 100,000 training steps. For the semi-supervised learning pipeline, both student and teacher are initialized with the respective baseline model and optimized for another 50,000 steps. For the student, the SGD optimizer is choosen with nesterov momentum and cosine decay learning rate scheduler [28, 7]. The pipeline is implement in pytorch and torchvision [29, 27] and run on NVIDIA V100 and A40 GPUs.

3.3. Post-Processing

Topographic maps are constructed by subtracting the softmax value of the interior class from one and a seed threshold is used to identify basins in there. These basins serve as the seed-points for an off-the-shelf watershed transform [8], which grows the basins from the seedpoints until a foreground threshold is reached [19]. Both, seed and foreground threshold, are optimized via grid-search on the validation dataset to yield the highest AP$_{.50}$ score. Then, the optimal thresholds are used for processing the test set predictions and performance metrics are calculated against the test set groundtruth.

4. Results

Baseline and semi-supervised trained student models are evaluated on three different datasets, comprised of cells with high variation in appearance, to show the generic applicability of our method.

- **DSB**: This dataset was published for the Kaggle 2018 Data Science Bowl competition. It consists of brightfield and fluorescence microscopy images of nuclei acquired under different conditions from different cells and magnifications. The dataset was split into 3800 training and 670 validation tiles of size 128×128 and 50 test set samples of various sizes.
- **Fly Wing**: This dataset consists of microscopy images of membrane labeled fly wings from the fruitfly

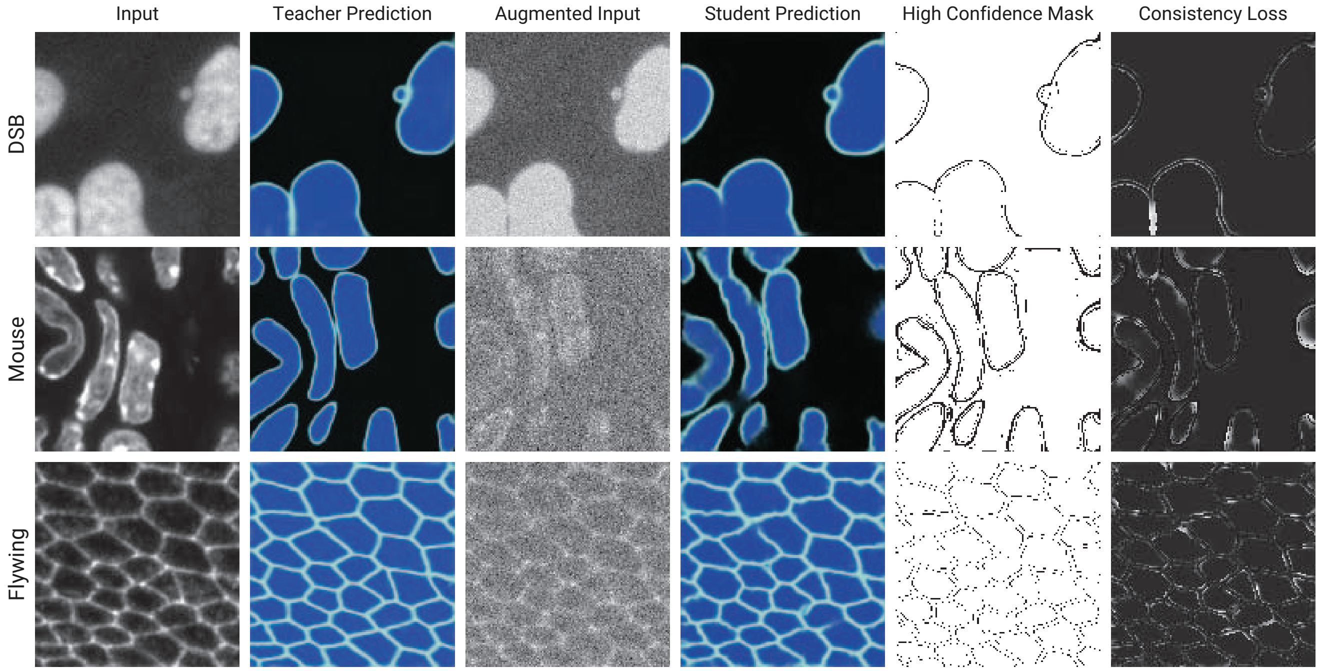

Figure 2. The figure shows the different transformations of an unlabeled sample that are used for augmentation consistency training. The teacher input data and prediction are shown in the left two columns, the augmented sample and the resulting student prediction are shown in the middle two columns and the resulting confidence-based loss filtering mask and the per-pixel loss L_{semi} is shown in the right two columns.

D. Melanogaster and consists of 1428 training and 252 validation tiles of 128×128 pixels and a test set of 50 images of size 512×512.

- **Mouse Nuclei**: This dataset consists of microscopy images of nuclei in the mouse skull which form clusters and have more diverse shapes than the nuclei in the other datasets. It consists of 908 training tiles and 160 validation tiles of size 128×128 pixels and a test set of 67 images of size 256×256.

All of the above datasets, tiled and split into subsets, were made publicly available by the authors of [4]. We found that multiple nuclei within individual tiles shared the same instance ID in the Mouse dataset. After clarifying with the authors [4], we decided to fix this by applying the connected components algorithm (i.e. `skimage.measure.label` with 2-hop connectivity [43]) on the labels and re-labeled the overlapping instance IDs.

Models are evaluated on the following scores choosen for benchmarking our approach with other published results. The number in the subscript denotes the Intersection over Union (IoU) threshold.

$$\text{AP}_{.50} = \frac{\text{TP}_{.50}}{\text{TP}_{.50} + \text{FP}_{.50}}$$

$$\text{AP}_{.10:.90} = \frac{1}{9} \cdot \sum_{i=.10}^{.90} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}$$

$$\text{S}_{DSB} = \frac{\text{TP}_{.50}}{\text{TP}_{.50} + \text{FP}_{.50} + \text{FN}_{.50}}$$

$$\text{F1}_{.50} = 2 \cdot \frac{\text{precision}_{.50} \cdot \text{recall}_{.50}}{\text{precision}_{.50} + \text{recall}_{.50}}$$

The $\text{AP}_{.10:.90}$ is averaged over different IoU thresholds, incremented by .10. All reported scores are calculated by averaging over the scores of individual test set samples.

This section is organized as follows: First a qualitative evaluation of the teacher and student predictions is conducted in 4.1, then the fully-supervised baseline model is compared with the semi-supervised model in 4.2 and finally our approach is compared to published results in 4.3.

4.1. Qualitative Evaluation

Figure 2 shows the different data representations used for the self-supervised augmentation consistency training. The pseudo labels for these samples are of high quality, whereas the student input is heavily augmented, resulting in local errors in the student predictions. The confidence-based mask,

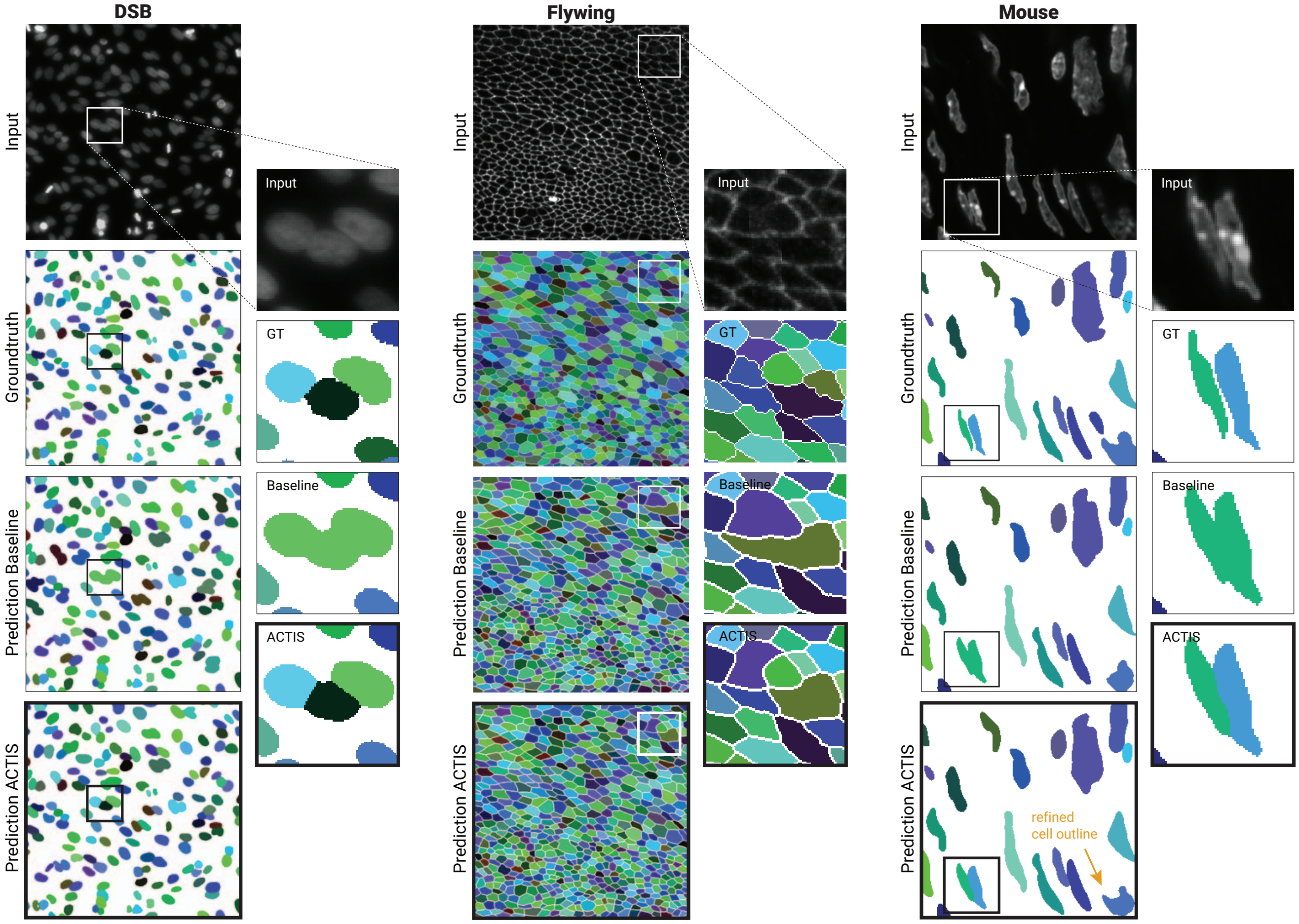

Figure 3. **Comparison to supervised baseline:** For all three datasets one example from the test set is displayed. The final (color-coded) instance predictions of our pipeline ACTIS are shown together with the results of the baseline and the groundtruth (GT). The insets highlight cases were ACTIS improves the segmentation of the baseline by splitting false merges.

calculated based on the pseudo-label, includes areas that contain inconsistencies in the student prediction and thus the loss at these regions is the highest as depicted by high intensity values in the consistency loss image. Visual inspection of the input and the student prediction confirms, that these areas contain segmentation errors and thus the resulting gradient of the loss is informative for improving the student.

Figure 3 presents test set examples from each dataset and respective predictions from the fully-supervised baseline model and ACTIS trained on 19 labeled samples each. ACTIS consistently reduces the number of false merges and thus improves overall performance. In single cases we have observed that ACTIS improves the cell shape as well (see Figure 3 orange arrow).

4.2. Comparison to the supervised baseline

Benchmark metrics for comparison of the fully-supervised baseline model and the semi-supervised model are presented in Table 1. Our approach consistently outperforms the baseline for various setups with different numbers of training samples. The only exception is the Flywing setup with $|D^l| = 76$, where the baseline model is slightly better than the semi-supervised model in terms of $AP_{.50}$, but their performance difference lies within one standard deviation of both models. Overall, the difference in performance between baseline and our approach is the highest for the more challenging DSB dataset, where scores are lower than for other datasets.

4.3. Benchmark against other approaches

In Table 2, the performance of our approach is compared to Denoiseg [4]. We use the same train/test/validation split as

Table 1. Test dataset scores, for the fully-supervised baseline model on the left and the semi-supervised ACTIS on the right, show that the latter consistently outperforms the baseline for various numbers of labeled samples $|D_l|$. All scores are averaged over 3 runs trained with a different sample of labeled images and the standard deviation is reported. Highest $AP_{.50}$ scores for comparison of baseline with ACTIS are highlighted in blue.

		Baseline	ACTIS		
	$\|D^l\|$	AP$_{.50}$	AP$_{.50}$	AP$_{.10:.90}$	F1$_{.50}$
DSB	10	.728±.030	.777±.010	.599±.019	.872±.006
	19	.782±.015	.801±.014	.604±.021	.889±.009
	38	.772±.053	.802±.024	.627±.023	.889±.016
	76	.798±.009	.821±.001	.636±.019	.901±.001
	152	.798±.022	.817±.009	.633±.027	.898±.006
	All	-	.847±.010	.650±.010	.916±.006
Flywing	5	.953±.002	.964±.001	.769±.015	.982±.001
	10	.955±.003	.965±.003	.767±.004	.982±.002
	19	.961±.004	.965±.000	.763±.008	.982±.000
	38	.966±.003	.968±.001	.776±.004	.984±.000
	76	.970±.002	.969±.001	.782±.003	.984±.001
	All	-	.972±.000	.790±.010	.986±.000
Mouse	5	.793±.003	.800±.003	.601±.018	.889±.002
	10	.794±.037	.814±.044	.615±.037	.897±.023
	19	.817±.002	.822±.007	.630±.017	.903±.004
	38	.843±.004	.850±.003	.657±.011	.919±.002
	76	.857±.011	.863±.009	.656±.010	.926±.005
	All	-	.872±.024	.678±.039	.927±.021

the authors, so that the comparison is fair. Denoiseg scores for the Mouse Nuclei dataset are higher in Table 2 than the ones reported in the publication [4], because we re-ran their code with 5 different seeds after we found and fixed a number of errors in the groundtruth annotations of this dataset. Our approach consistently outperforms Denoiseg by a large margin over all conducted experiments. The reasons for this are, that Denoiseg uses a Vanilla U-Net [32], without a pretrained encoder, which has a large effect on the performance as we show in our ablation study in Table 3. In addition, it only uses a very basic augmentation pipeline, consisting of 90-degree rotations, flips and additive gaussian noise. In addition, Denoiseg adds self-supervised denoising as an auxiliary task and adds the denoised image as an additional output domain, whereas ACTIS does denoising (and other tasks such a normalization and sharpening) directly on the 3-class domain that is later used for segmentation.

4.4. Ablation study

Table 3 reports results for an ablation study where single components of ACTIS were ablated. The first row reports the results of the full pipeline with 19 labeled samples for each of the datasets. In the second row, the momentum teacher is replaced by a fixed student model (Momentum Teacher ✗) and the performance drops slightly for all datasets. If confidence filtering is dropped and the student is trained on the quantized teacher predictions only, the performance degrades substantially for mouse and DSB, whereas performance on the flywing dataset only slightly decreases. This is due to the substantially lower number of segmentation errors in the predictions of the flywing dataset and therefore the lower number of errors that would need to be filtered out. The second to last row is the baseline model, where models are trained fully-supervised on a small number of samples. In the last row, imagenet pretrained weights for the encoder are not used and as a result the performance declines considerably across the board. Despite the domain gap from imagenet to the bio-medical datasets used in this paper, the features learned seem to transfer well.

Table 2. Test dataset S_{DSB} scores at IoU=0.5 for Denoiseg (left) and ACTIS (right) show that our approach consistently outperforms the other (highlighted in blue). Scores for Denoiseg are averaged over 5 different labeled/un-labeled splits, whereas results for ACTIS are averaged over 3 splits and the standard deviations are reported.

		Denoiseg[4]	ACTIS
Dataset	$\|D^l\|$	S_{DSB}	S_{DSB}
DSB	10	.690±.006	.791±.010
	19	.705±.005	.813±.012
	38	.718±.004	.813±.023
	76	.728±.005	.831±.001
	152	.757±.003	.828±.008
Flywing	5	.882±.014	.964±.001
	10	.907±.003	.965±.003
	19	.899±.005	.965±.000
	38	.923±.003	.968±.001
	76	.929±.001	.969±.008
Mouse	5	.721±.014	.808±.003
	10	.730±.022	.823±.041
	19	.755±.013	.830±.007
	38	.768±.013	.856±.003
	76	.795±.010	.868±.009

5. Discussion

The results show that semi-supervised augmentation consistency training can substantially improve data efficiency. For the DSB and mouse dataset, the effect of using ACTIS over the fully-supervised baseline is for some setups comparable to more than doubling the number of labeled samples used for training according to Table 1. Results also indicate that ACTIS is especially beneficial for models trained on the DSB dataset, which, unlike the other datasets, consists of samples from different imaging modalities and experimental conditions. Therefore, it might be that our approach works especially well for hetereogenous datasets where part of the problem is adapting to the different domains contained in the dataset. This is in line with the literature on unsupervised domain adaptation for semantic segmentation,

Table 3. Different components of the pipeline are ablated and the $\text{AP}_{.50}$ score of the respective models on the test dataset is reported for setups trained with 19 labeled samples.

Ablations				$\text{AP}_{.50}$ split by dataset		
Include	Momentum	Confidence	Pre-trained			
L_{semi}	Teacher	Filtering	Encoder	DSB	Mouse	Flywing
✓	✓	✓	✓	.801±.014	.822±.007	.965±.000
✓	✗	✓	✓	.800±.011	.805±.021	.964±.000
✓	✓	✗	✓	.786±.021	.797±.020	.965±.000
✗	✗	✗	✓	.782±..015	.817±.002	.961±.004
✗	✗	✗	✗	.715±.010	.795±.009	.955±.000

where similar augmentation consistency approaches were already successfully applied [31, 1].

Composing the data augmentation pipeline and setting reasonable hyperparameters is key for making augmentation consistency training work and improve performance over the baseline. More sophisticated data augmentation methods like mixup [49], cutout [11] or copy-paste augmentation [13] could further improve results, but were out of scope for this work. Also, poisson noise could be used instead of gaussian noise for augmentation, since it better resembles real noise distributions observed in microscopy images [22]. In addition, more work is necessary to clarify which augmentations to use for color images in the bio-medical domain such as Hematoxilin and Eosin stained whole slide tissue images.

Better performing instance segmentation models such as StarDist [36], Embedseg [24] or the 3-class model with offset vectors as auxiliary task [19] could be used instead of the vanilla 3-class model [5] to further boost performance.

A key assumption for the confidence filtering step in ACTIS is, that the confidence scores predicted by the teacher model are well calibrated, meaning that they are lower for regions that contain more errors and vice versa. However, this assumption might be violated to a certain degree and our U-Net may be overconfident in its predictions [16]. Therefore, more sophisticated and better calibrated uncertainty quantification approaches such as approximate bayesian methods [12, 34] could enable further improvements.

Despite showing considerable improvement in performance (c.f. Table 3), the imagenet-pre-trained encoder in our U-Net architecture is probably not the optimal initialization for our models. Self-supervised generative pre-training, on the whole dataset or on a larger array of different datasets from the bio-medical domain, might provide better initializations and could further improve performance [42].

6. Limitations

Despite using only a small number of labeled training set tiles, the number of labeled tiles in the validation set for all experiments was held constant at 15% of the sum of tiles in the training and validation dataset. Thus, when accounting for the validation dataset, the proposed method is less label efficient than anticipated. This problem also applies to other approaches, which use a similar percentage of labeled data for validation [46, 4]. Therefore, more data efficient ways to estimate the generalization capabilities of models during training are highly desired to further decrease the required number of labeled samples.

7. Conclusion

We presented a simple and highly performant approach for semi-supervised instance segmentation. The pipeline, consisting of a pre-trained encoder, coupled with standard augmentations, a simple confidence based loss-masking scheme for consistency regularization and a momentum-updated teacher model is easy to implement and should be applicable to other datasets. In addition, the approach should be applicable to other dense prediction tasks in the bio-medical domain such as semantic or panoptic segmentation which we will explore in the future.

Acknowledgements We wish to thank the anonymous reviewers for their helpful comments. Funding: J.L.R. and J.F.: German Research Foundation RTG 2424. J.L.R, J.F., P.H., J.P.A., and D.K.: Max-Delbrueck-Center for Molecular Medicine in the Helmholtz Association (MDC) and Helmholtz Imaging. J.L.R.: IFI program of the German Academic Exchange Service (DAAD). P.H.: MDC-NYU exchange program and HFSP grant RGP0021/2018-102.

References

[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15384–15394, 2021.

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022.

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

[4] Tim-Oliver Buchholz, Mangal Prakash, Deborah Schmidt, Alexander Krull, and Florian Jug. Denoiseg: joint denoising and segmentation. In *European Conference on Computer Vision*, pages 324–337. Springer, 2020.

[5] Juan C Caicedo, Jonathan Roth, Allen Goodman, Tim Becker, Kyle W Karhohs, Matthieu Broisin, Csaba Molnar, Claire McQuin, Shantanu Singh, Fabian J Theis, et al. Evaluation of deep learning strategies for nucleus segmentation in fluorescence images. *Cytometry Part A*, 95(9):952–965, 2019.

[6] Long Chen, Weiwen Zhang, Yuli Wu, Martin Strauch, and Dorit Merhof. Semi-supervised instance segmentation with a learned shape prior. In *International Workshop on Interpretability of Machine Intelligence in Medical Image Computing*, pages 94–102. Springer, 2020.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[8] Luis Pedro Coelho. Mahotas: Open source software for scriptable computer vision. *arXiv preprint arXiv:1211.4907*, 2012.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W Remedios, Shunxing Bao, Bennett A Landman, Lee E Wheless, Lori A Coburn, Keith T Wilson, et al. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*, 2023.

[11] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[13] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021.

[14] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

[15] Noah F Greenwald, Geneva Miller, Erick Moen, Alex Kong, Adam Kagel, Thomas Dougherty, Christine Camacho Fullaway, Brianna J McIntosh, Ke Xuan Leow, Morgan Sarah Schwartz, et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature biotechnology*, 40(4):555–565, 2022.

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[18] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019.

[19] Peter Hirsch and Dagmar Kainmueller. An auxiliary task for learning nuclei segmentation in 3d microscopy images. In *Medical Imaging with Deep Learning*, pages 304–321. PMLR, 2020.

[20] Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P Lungren, Serena Yeung, and Akshay S Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digital Medicine*, 6(1):74, 2023.

[21] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

[22] Alexander Krull, Hector Basevi, Benjamin Salmon, Andre Zeug, Franziska Müller, Samuel Tonks, Leela Muppala, and Ales Leonardis. Image denoising and the generative accumulation of photons. *arXiv preprint arXiv:2307.06607*, 2023.

[23] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[24] Manan Lalit, Pavel Tomancak, and Florian Jug. Embedseg: Embedding-based instance segmentation for biomedical microscopy data. *Medical image analysis*, 81:102523, 2022.

[25] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013.

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[27] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1485–1488, 2010.

[28] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/k2). In *Dokl. Akad. Nauk. SSSR*, volume 269, page 543, 1983.

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[30] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

[31] Viraj Uday Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Augmentation consistency-guided self-training

for source-free domain adaptive semantic segmentation. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[33] Josef Lorenz Rumberger, Elias Baumann, Peter Hirsch, Andrew Janowczyk, Inti Zlobec, and Dagmar Kainmueller. Panoptic segmentation with highly imbalanced semantic labels. In *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*, pages 1–4. IEEE, 2022.

[34] Josef Lorenz Rumberger, Lisa Mais, and Dagmar Kainmueller. Probabilistic deep learning for instance segmentation. In *European Conference on Computer Vision*, pages 445–457. Springer, 2020.

[35] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016.

[36] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 265–273. Springer, 2018.

[37] Andrew B Sellergren, Christina Chen, Zaid Nabulsi, Yuanzhen Li, Aaron Maschinot, Aaron Sarna, Jenny Huang, Charles Lau, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, et al. Simplified transfer learning for chest radiography models using less data. *Radiology*, 305(2):454–465, 2022.

[38] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[39] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021.

[40] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[41] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[42] Keyu Tian, Yi Jiang, Qishuai Diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing bert for convolutional networks: Sparse and hierarchical masked modeling. *arXiv preprint arXiv:2301.03580*, 2023.

[43] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[44] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022.

[45] Zhenyu Wang, Yali Li, and Shengjin Wang. Noisy boundaries: Lemon or lemonade for semi-supervised instance segmentation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16826–16835, 2022.

[46] Huisi Wu, Zhaoze Wang, Youyi Song, Lin Yang, and Jing Qin. Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11666–11675, 2022.

[47] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

[48] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[49] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.